National Statistical Service

Confidentiality Information Sheet 3

Confidentiality, and therefore trust, can be broken when a person or an organisation can be identified in a disseminated dataset, either directly or indirectly.

Confidentiality

Managing identification risks

One of the biggest challenges in making data publicly available is ensuring that no person or organisation is likely to be identified in the data.

Identification (often referred to as disclosure) occurs when someone learns something that they did not already know about another person or organisation through data that has been disseminated. This may be in the form of aggregate data (typically data presented in tables) or microdata (unit record data where each record represents observations for a person or an organisation).

Managing the risk of identification (disclosure control)

The process for managing the risk of identification is the same for any data that are disseminated, whether it is aggregate data or microdata. The first step is to assess potential identification risks. The second step is to manage the risks of identification by using an appropriate method to confidentialise the data.

The key elements in the process of managing the risk of identification are outlined in the diagram opposite.

A process for managing identification risk

Understand the legal and ethical obligations to protect the confidentiality of data
For more detail see Information sheets 1 and 2

↓

Establish policies and procedures to meet confidentiality obligations
For more detail see Information sheets 3, 4 and 5

↓

De-identify the data (remove any direct identifiers from the dataset)

↓

Assess *potential identification* risks by evaluating the factors that contribute to the likelihood of identification
For more detail see Information sheets 3, 4 and 5

↕ Test and evaluate to mitigate risk

Manage the risks of identification – confidentialise
For more detail see Information sheets 3, 4 and 5

↓

Provide safe access to data
For more detail see Information sheet 5

Definitions

Microdata are unit record data containing individual responses to questions on survey questionnaires, or administrative forms. For example, data in response to the question 'In what year were you born?'.

Aggregate data (or macrodata) refers to aggregated microdata. For example, a count of the number of people of a particular age (obtained from the question 'In what year were you born?').

For more information about managing identification risks, or to provide feedback on this series, please email: inquiries@nss.gov.au

Confidentiality – managing identification risks

Assessment of potential identification risks

The likelihood of a person or an organisation being identified, and the confidentiality methods used to minimise the risk of identification, will vary depending on a range of factors including:

- the amount of detail included in the data (the more detail, the higher the risk of identification);
- the sensitivity of the variables within the data (data items such as financial, health and medical or criminal record information may increase the risk of identification and would cause serious public concern if disclosed); and
- how the information is presented (aggregate data released in tables poses less risk of identification than the release of microdata).

Confidentialisation to manage potential identification risk

The objective of confidentialisation is to meet legal and ethical obligations to protect the identity and privacy of individuals and organisations, while at the same time maximising the usefulness of the data for statistical and research purposes. While confidentialisation can be used to manage the risk of identification, it may not completely eliminate the risk.

There are two general methods (often referred to as statistical disclosure control methods) used to confidentialise data that are to be disseminated:

- data modification methods (perturbation) which involve changing the data slightly to reduce the risk of disclosure, while retaining as much content and structure as possible; and
- data reduction methods which aim to control or limit the amount of detail available, without compromising the overall usefulness of the information available for research.

The techniques used to confidentialise data are similar for both aggregate data and microdata. However, the way in which the risks are assessed and the confidentiality techniques which are applied are different:

- for aggregate data, the techniques are applied to cells in a table identified as unsafe, after aggregation; and
- for microdata, the techniques are applied to data items within individual unit records, or to individual unit records, identified as unsafe before aggregation or analysis.

This information sheet provides a broad overview of the process involved in managing the risks of identification. More detailed information is given in the subsequent information sheets in this series.

Information Sheet 4: 'How to confidentialise data: the basic principles' focuses on the risk assessment that applies to aggregate data presented in tables, and the popular confidentiality techniques that apply to both aggregate and microdata.

Information Sheet 5: 'Managing the risk of disclosure in the release of microdata' focuses on the risk assessment and confidentiality considerations that apply to microdata.